Joint Image and Depth Estimation with Mask-Based Lensless Cameras

Yucheng Zheng and M. Salman Asif

***Abstract*—Mask-based lensless cameras replace the lens of a conventional camera with a customized mask. These cameras can potentially be very thin and even flexible. Recently, it has been demonstrated that such mask-based cameras can recover light intensity and depth information of a scene. Existing depth recovery algorithms either assume that the scene consists of a small number of depth planes or solve a sparse recovery problem over a large 3D volume. Both these approaches fail to recover scene with large depth variations. In this paper, we propose a new approach for depth estimation based on alternating gradient descent algorithm that jointly estimates a continuous depth map and light distribution of the unknown scene from its lensless measurements. The computational complexity of the algorithm scales linearly with the spatial dimension of the imaging system. We present simulation results on image and depth reconstruction for a variety of 3D test scenes. A comparison between the proposed algorithm and other method shows that our algorithm is faster and more robust for natural scenes with a large range of depths.**

***Index Terms*—Lensless imaging, flatcam, depth estimation, non-convex optimization, alternating minimization.**

I. Introduction

Depth estimation is an important and challenging problem that arises in a variety of applications including computer vision, robotics, and autonomous systems. Existing depth estimation systems use stereo pairs of conventional (lens-based) cameras or time-of-flight sensors [1]–[3]. These cameras can be heavy, bulky and require large space for their installation. Therefore, their adoption for portable and lightweight devices with strict physical constraints is still limited.

In this paper, we propose a joint image and depth estimation framework for a computational lensless camera that consists of a fixed, binary mask placed on top of a bare sensor. Such mask-based cameras offer an alternative design for building cameras without lenses. A recent example of mask-based lensless camera is known as FlatCam [4]. In contrast with a lens-based camera that is designed to map every point in the scene to a single pixel on the sensor, every sensor in a FlatCam records light from every point in the scene. A single point source in the scene casts a shadow of the mask on the sensor, which shifts if the point moves parallel to the sensor plane and expand/shrink if the point source moves toward/away from the sensor plane. The measurements recorded on the sensor thus represent superposition of shifted and scaled versions of the mask shadows corresponding to light sources in different directions and depths. Image and depth information about the scene is thus encoded in the measurements, and we can solve an inverse problem to estimate both of them.

Joint estimation of intensity and depth is a nonconvex problem. In fact, estimation of depth by itself, even with known scene intensity, is a nonconvex problem. To jointly estimate depth and light distribution, we propose a two step approach that consists of an initialization step and an alternating gradient descent step to minimize our objective. To preserve sharp edges in the image intensity and depth map, we include an adaptive regularization penalty in our objective function. An overview of the reconstruction framework is illustrated in Figure 1. We initialize the estimates of image intensity and depth using a greedy algorithm proposed in [5]. Then we refine the estimates by minimizing an objective function with respect to image intensity and depth via alternating gradient descent. To simplify the recovery algorithm, we assume that the mask pattern is differentiable everywhere. We use adaptive weights to add smoothness regularization on the intensity and depth estimates [6]. Even though the problem of joint estimation of intensity and depth is non-convex, we observed that a simple regularization makes the algorithm robust against local minima of the loss function and improve the performance of the algorithm.

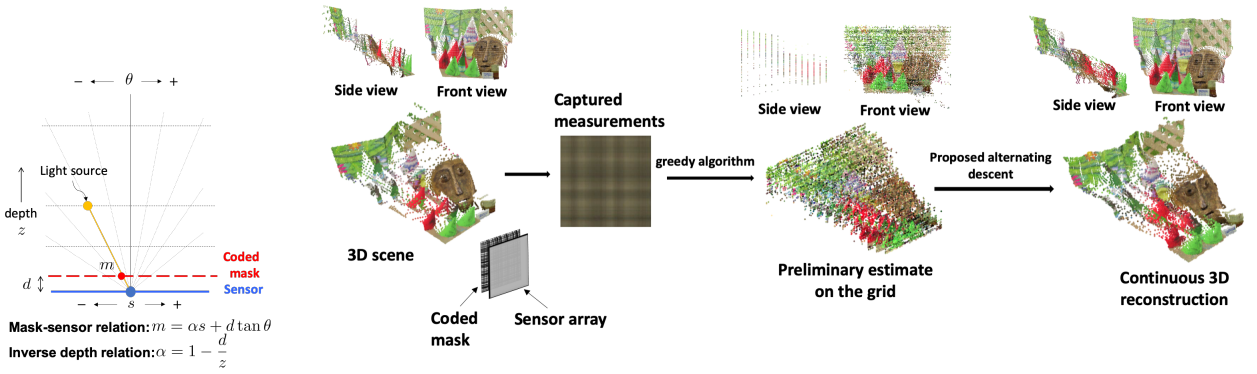
The key contributions of this paper are as follows.

- We propose a new computational framework for joint estimation of light intensity and depth map from a single image of a mask-based lensless camera. In contrast to other methods, our method estimates the depth map on a continuous domain. Our algorithm consists of a careful initialization step based on greedy pursuit and an alternating minimization step based on gradient descent.
- The problem of joint image and depth recovery is highly nonconvex. To tackle this issue, we present different regularization schemes that offer robust recovery on a diverse dataset.
- We present simulation results on standard 3D datasets and demonstrated a significant improvement over existing methods for 3D imaging using coded mask-based lensless cameras.

II. Related Work

A pinhole camera, also known as *camera obscura*, is the simplest example of a mask-based lensless camera. Even though a pinhole can easily provide an image of the scene onto a sensor plane, the image quality is often severely affected by noise because the amount of light collected is limited by the pinhole aperture [7]. Coded aperture-based lensless

Y. Zheng and M. Asif are with the Department of Electrical and Computer Engineering, University of California, Riverside, CA, 92521 USA (e-mail: yzhen069@ucr.edu; sasif@ece.ucr.edu).



(a) 1D imaging model for a planar sensor with a coded mask placed at distance d. Light rays from a light source at location (θ, z) are received by all the sensor pixels. A light ray that hits sensor pixel s passes through mask at location m.

(b) An overview of the proposed intensity and depth estimation framework. Consider a natural scene as a 3D point cloud, where each point represents a light source located at a different depth. The camera consists of a fixed, coded mask placed on top of an image sensor. Every point in the scene casts a shadow of the mask on the sensor plane. Each sensor pixels records a linear combination of the scene modulated by the mask pattern. The recovery algorithm consists of two steps. (1) Initialization using a greedy depth selection method. (2) An alternating gradient descent-based refinement algorithm that jointly estimates the light distribution and depth map on a continuous domain.

Fig. 1: A coded mask-based imaging model and an overview of the proposed continuous depth estimation framework.

cameras avoid this problem by increasing the number of pinholes and allowing more light to reach the sensor [4], [8]–[11]. In contrast to a pinhole camera where only one inverted image of the scene is obtained through a single pinhole, the measurements captured through a coded-mask are a linear combination of all the pinhole images under every mask element. To recover an image of the scene, we need to solve a computational image recovery problem [4], [8].

A coded aperture system offers another advantage by encoding light from different directions and depths differently. Note that a bare sensor can provide the intensity of a light source but not its spatial location. A mask in front of the sensor encodes directional information of the source in the sensor measurements. Consider a single light source with a dark background; the image formed on the sensor will be a shadow of the mask. If we change the angle of the light source, the mask shadow on the sensor will shift. Furthermore, if we increase or decrease the depth of the light source, the width of the shadow will decrease or increase, respectively. Thus, we can represent the relationship between all the points in the 3D world and the sensor measurements as a linear system, which depends on the pattern and the placement of the mask. We can solve this system using an appropriate computational algorithm to recover the image of the scene.

The depth-dependent imaging capability in coded aperture systems is known since the pioneering work in this domain [8], [12]. The following excerpt in [8] summarizes it well: *"One can reconstruct a particular depth in the object by treating the picture as if it was formed by an aperture scaled to the size of the shadow produced by the depth under consideration."* However, the classical methods usually assume that the scene consists of a single plane at known depth. In this paper, we assume that the depth map is arbitrarily distributed on a continuous domain and the true depth map is unknown at the time of reconstruction.

The 3D lensless imaging problem has also recently been studied in [5], [11], [13], [14]. These methods can broadly be divided into two categories. In the first category, the 3D scene is divided into a finite number of voxels. To recover the 3D light distribution, these methods solve an ℓ_1 norm-based recovery problem under the assumption that the scene is very sparse [13], [14]. In the second category, the 3D scene is divided into an intensity map and multiple depth planes such that each pixel is assigned one intensity and depth. To solve the intensity and depth recovery problem, these methods either sweep through the depth planes [11] or assign depth to each pixel using a greedy method [5]. Our proposed method belongs to the second category in which we model the image intensity and depth separately and assume that the depth values of the scene are distributed on a continuous domain. To recover 3D scene, we jointly estimates image intensity and depth map from the available sensor measurements.

Joint estimation of image intensity and depth map can be viewed as a nonlinear inverse problem in which the sampling function is dependent on scene depth. Similar inverse problem also arises in many other fields such as direction-of-arrival estimation in radar [15], super resolution [16] and compressed sensing [17]–[19]. Similar to joint estimation of image intensity and depth, the solution approaches to these problems consists of two main steps: identification of signal bases and the estimation of signal intensities based on the identified bases. The problem of identifying the signal bases from continuously varying candidates is often called off-the-grid signal recovery. The methods for solving the off-the-grid signal recovery problems can be divided into two main types. The first approach formulates the problem as a convex program on a continuous domain and solve it using an atomic norm minimization approach [20], [21]. The second approach linearizes the problem w.r.t. the continuous optimization parameter using a first-order approximation at every iteration [16], [22]. Our proposed algorithm is inspired by the second approach.

Mask-based lensless cameras have traditionally been used for imaging light at wavelengths beyond the visible spectrum [9], [10]. Other examples related to mask-based cameras include controllable aperture and employing coded-mask for compressed sensing and computational imaging [23], [24], single pixel camera [25] and external mask setting [26].

Coded masks have also recently been used with conventional lens-based cameras to estimate depth and lightfield [27], [28]. Recently, a number of data-driven methods have been proposed to design custom phase masks and optical elements to estimate depth from a single image [29], [30]. An all-optical diffractive deep neural network is proposed in [31], [32], which can perform pattern recognition tasks such as handwritten digits classification using optical mask layers. Such networks can literally process images at a lightning-fast pace with near-zero energy cost.

III. Methods

A. Imaging Model

We divide the 3D scene under observation into $N \times N$ uniformly spaced directions. We use θ_i and θ_j to denote the angular directions of a light source with respect to the center of the sensor. The intensity and depth of the light source are denoted using $l_{i,j}$ and $z_{i,j}$ respectively. Figure 1(a) depicts the geometry of such an imaging model. A planar coded-mask is placed on top of a planar sensor array at distance d. The $M \times M$ sensor array captures lights coming from the scene modulated by the coded-mask.

Every light source in the scene casts a shadow of the mask on the sensor array, which we denote using basis functions ψ. We use s_u and s_v to index a pixel on the rectangular sensor array. The shadow cast by a light source with unit intensity at $(\theta_i, \theta_j, z_{i,j})$ can be represented as the following basis or point spread function:

$$\psi_{i,j}(s_u, s_v) = \text{mask}\ [\alpha_{i,j} s_u + d\tan(\theta_i), \alpha_{i,j} s_v + d\tan(\theta_j)], \tag{1}$$

where mask$[u, v]$ denotes the transmittance of the mask pattern at location (u, v) on the mask plane and $\alpha_{i,j}$ is a variable that is related to the physical depth $z_{i,j}$ with the following inverse relation:

$$\alpha_{i,j} = 1 - \frac{d}{z_{i,j}}, \tag{2}$$

If the 3D scene consists of only a single point source at (θ_i, θ_j) with light intensity $l_{i,j}$, the measurement captured at sensor pixel (s_u, s_v) would be

$$y(s_u, s_v) = \psi_{i,j}(s_u, s_v) l_{i,j}. \tag{3}$$

The measurement recorded on any sensor pixel is the summation of contributions from each of the point sources in the 3D scene. The imaging model for a single sensor pixel can be represented by

$$y(s_u, s_v) = \sum_{i=1}^{N} \sum_{j=1}^{N} \psi_{i,j}(s_u, s_v) l_{i,j}. \tag{4}$$

We can write the imaging model for all the sensors in a compact form as

$$\mathbf{y} = \mathbf{\Psi}(\boldsymbol{\alpha})\mathbf{l} + e, \tag{5}$$

where $\mathbf{y} \in \mathbb{R}^{M^2}$ is a vectorized form of an $M \times M$ matrix that denotes sensor measurements, $\mathbf{l} \in \mathbb{R}^{N^2}$ is a vectorized form of an $N \times N$ matrix that denotes light intensity from all the locations $(\theta_i, \theta_j, \alpha_{i,j})$, and $\mathbf{\Psi}$ is a matrix with all the basis functions corresponding to $\theta_i, \theta_j, \alpha_{i,j}$. The basis functions in (5) are parameterized by the unknown $\boldsymbol{\alpha} \in \mathbb{R}^{N^2}$. e denotes noise and other nonidealities in the system.

We can jointly estimate light distribution ($\mathbf{l}$) and inverse depth map ($\boldsymbol{\alpha}$)[1] using the following optimization problem:

$$\underset{\boldsymbol{\alpha}, \mathbf{l}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \mathbf{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2. \tag{6}$$

Note that if we know the true values of $\boldsymbol{\alpha}$ (or we fix it to something), then the problem in (6) reduces to a linear least-squares problem that can be efficiently solved via standard solvers. On the other hand, if we fix the value of $\mathbf{l}$, the problem remains nonlinear with respect to $\boldsymbol{\alpha}$. In the next few sections we discuss our approach for solving the problem in (6) via alternating minimization.

B. Initialization

Since the minimization problem in (6) is not convex, a proper initialization is often needed to ensure convergence to a local minima close to the optimal point. A naïve approach is to initialize all the point sources in the scene at the same depth plane. To select an initial depth plane, we sweep through a set of candidate depth planes and perform image reconstruction on one depth plane at a time by solving the following linear least squares problem:

$$\underset{\mathbf{l}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \mathbf{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2. \tag{7}$$

We evaluate the loss value for all the candidate depth planes and picked the one with the smallest loss as our initialized depth. The mask basis function in (1) changes as we change $\boldsymbol{\alpha}$, which has an inverse relation with the scene depth. We select candidate depth corresponding to uniformly sampled values of $\boldsymbol{\alpha}$, which yields nonuniform sampling of the physical scene depth. The single-depth initialization approach is computationally simple and provides a reasonable initialization of light distribution to start with, especially when the scene is far from the sensor.

Our second approach for initialization is the greedy method proposed in [5]. Greedy algorithms are widely used for sparse signal recovery [17]–[19]. Based on these algorithms, [5] proposed a greedy depth pursuit algorithm for depth estimation from FlatCam [4]. The algorithm works by iteratively updating the depth surface that matches the observed measurements the best.

The depth pursuit method assumes that the scene consists of a finite number of predefined depth planes. We start the program by initializing all the pixels at a single depth plane

[1] $\boldsymbol{\alpha}$ has an inverse relation with the depth map (2); therefore we refer to it as inverse depth map throughout the paper.

and the estimation of light intensities $\mathbf{l}$ based on initialized depth map. The first step is to select new candidate values for $\boldsymbol{\alpha}$. The new candidates are selected using the basis vectors that are mostly correlated with the current residual of the estimate. In the second step, new candidates for $\boldsymbol{\alpha}$ are appended to the current estimate. We solve a least squares problem using the appended $\boldsymbol{\alpha}$. In the third step, we prune the $\boldsymbol{\alpha}$ by selecting $\alpha_{i,j}$ as the value corresponding to the largest magnitude of $\mathbf{l}_{i,j}$. Although this method may not estimate the off-grid point sources well, it produces a good preliminary estimate of the scene.

C. Refinement via Alternating Gradient Descent

To solve the minimization problem in (6), we start with the preliminary image and depth estimates from the initialization step and alternately update depth and light distribution via gradient descent. The main computational task in gradient descent method is computing the gradient of the loss function w.r.t. $\boldsymbol{\alpha}$. To compute that gradient, we expand the loss function in (6) as

$$L = \frac{1}{2}\sum_{u,v=1}^{M}\left(y(s_u,s_v) - \sum_{i,j=1}^{N}\psi_{i,j}(s_u,s_v)l_{i,j}\right)^2 \tag{8}$$

We define $R_{u,v} = y(s_u,s_v) - \sum_{i,j=1}^{N}\psi_{i,j}(s_u,s_v)l_{i,j}$ as the residual approximation error at location (s_u,s_v). The derivatives of the loss function with respect to the $\alpha_{i,j}$ is given as

$$\begin{aligned}\frac{\partial L}{\partial \alpha_{i,j}} &= \sum_{u,v=1}^{M} R_{u,v}\frac{\partial R_{u,v}}{\partial \alpha_{i,j}} \\ &= -l_{i,j}\sum_{u,v=1}^{M} R_{u,v}\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial \alpha_{i,j}}.\end{aligned} \tag{9}$$

We compute the derivatives of sensor value with respect to the $\alpha_{i,j}$ using the total derivative [2] as follows.

$$\begin{aligned}\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial \alpha_{i,j}} &= \frac{\partial \psi_{i,j}(s_u,s_v)}{\partial u_{i,j}}\frac{\partial u_{i,j}}{\partial \alpha_{i,j}} + \frac{\partial \psi_{i,j}(s_u,s_v)}{\partial v_{i,j}}\frac{\partial v_{i,j}}{\partial \alpha_{i,j}} \\ &= \frac{\partial \psi_{i,j}(s_u,s_v)}{\partial u_{i,j}}s_u + \frac{\partial \psi_{i,j}(s_u,s_v)}{\partial v_{i,j}}s_v.\end{aligned} \tag{10}$$

$u_{i,j} = \alpha_{i,j}s_u + d\tan(\theta_i)$ and $v_{i,j} = \alpha_{i,j}s_v + d\tan(\theta_j)$ denote two dummy variables that also correspond to the specific location on the mask where a light ray from a point source at angle (θ_i,θ_j) and depth $\alpha_{i,j}$ and sensor pixel at (s_u,s_v) intersects with the mask plane. The terms in $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial u_{i,j}}$, $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial v_{i,j}}$ can be viewed as the derivatives of mask pattern along the respective spatial coordinates and evaluated at $u_{i,j}, v_{i,j}$. We compute these derivatives using finite-difference of $\psi_{i,j}(s_u,s_v)$ over a fine grid and linear interpolation.

D. Algorithm Analysis

To solve the non-linear least squares problem (7) in our algorithms, we compute the gradient derived in (10) and use it as input of a optimization solver. Suppose ψ_i and ψ_j denote the basis function vector evaluated on 1D mask,

$$\begin{aligned}\psi_i(s_u) &= \text{mask }[\alpha_{i,j}s_u + d\tan(\theta_i)] \\ \psi_j(s_v) &= \text{mask }[\alpha_{i,j}s_v + d\tan(\theta_j)],\end{aligned} \tag{11}$$

If we use a separable mask pattern that the mask pattern we use is the outer product of two 1D mask patterns, the 2D mask function $\psi_{i,j}$ in (1) can be computed as the outer product of two vectors given as $\psi_{i,j} = \psi_i\psi_j^T$. Similarly, we define 1D sub-gradient function g as

$$\begin{aligned}g_i(s_u) &= \frac{\partial \psi_{i,j}(s_u,s_v)}{\partial u_{i,j}} \\ g_j(s_v) &= \frac{\partial \psi_{i,j}(s_u,s_v)}{\partial v_{i,j}},\end{aligned} \tag{12}$$

Similar to (10), the functions $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial u_{i,j}}$ and $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial v_{i,j}}$ are the sub-gradient functions along the 1D mask. It takes non-negative values at locations where mask pattern value changes and takes zero value at the other places. Using the derivation in (10), the matrix contains $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial \alpha_{i,j}}$ at all (s_u,s_v) can be computed using the following sum of two vector outer products.

$$\frac{\partial \psi_{i,j}}{\partial \alpha_{i,j}} = g_i\psi_j^T + \psi_i g_j^T \tag{13}$$

Using the derivations in (9), the derivative of loss function with respect to depth value can be computed using the following matrix multiplications, where R refers to the matrix of residual $R_{u,v}$ at all (s_u,s_v)

$$\frac{\partial L}{\partial \alpha_{i,j}} = g_i^T R\psi_j + \psi_i^T R g_j \tag{14}$$

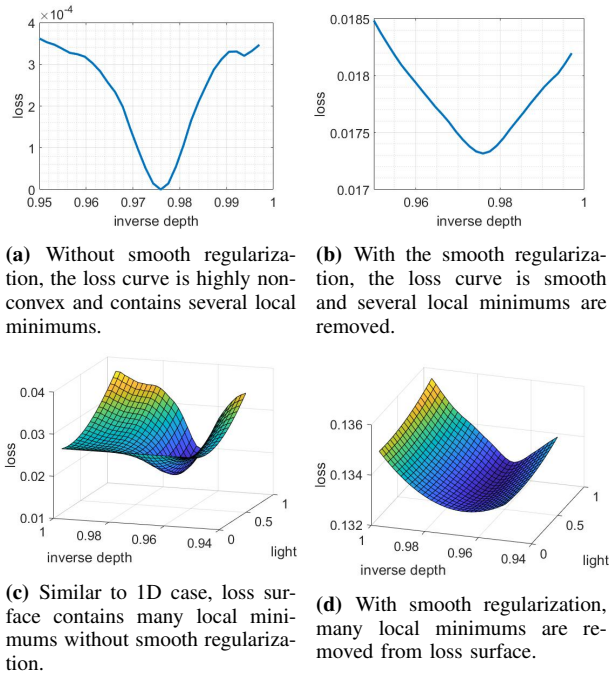
Suppose we have $M \times M$ pixels on sensor array. The computation in (14) takes $2M^2 + 2M$ multiplications. We then feed our gradients to `minfunc` solver [33] with L-BFGS algorithm [34] to solve the non-linear optimization problem in (7).

E. Regularization Approaches

ℓ_2 regularization on spatial gradients. The optimization problem in (6) is highly non-convex and contains several local minima; therefore, the estimate often gets stuck in some local minima and the estimated intensity and depth map are coarse. To improve the performance of our algorithm for solving the non-convex problem in (6), we seek to exploit additional structures in the scene. A standard assumption is that the depth of neighboring pixels is usually close, which implies that the spatial differences of (inverse) depth map are small. To incorporate this assumption in our model, we add a quadratic regularization term on the spatial gradients of the inverse depth map to our loss function. The quadratic regularization term is defined on an $N \times N$ inverse depth map matrix $\boldsymbol{\alpha}$ and can be written as

$$\begin{aligned}R(\boldsymbol{\alpha}) &= \sum_{i,j=1}^{N}(\alpha_{i,j} - \alpha_{i+1,j})^2 + (\alpha_{i,j} - \alpha_{i,j+1})^2 \\ &= \|\nabla_r\boldsymbol{\alpha}\|_F^2 + \|\nabla_c\boldsymbol{\alpha}\|_F^2,\end{aligned} \tag{15}$$

[2] Recall that the total derivative of a multivariate function $f(x,y)$ is $\frac{\partial f(x,y)}{\partial x}dx + \frac{\partial f(x,y)}{\partial y}dy$.



(a) Without smooth regularization, the loss curve is highly non-convex and contains several local minimums.

(b) With the smooth regularization, the loss curve is smooth and several local minimums are removed.

(c) Similar to 1D case, loss surface contains many local minimums without smooth regularization.

(d) With smooth regularization, many local minimums are removed from loss surface.

Fig. 2: A comparison between objective loss functions without and with smooth regularization. The inverse depth axis refers to the value of α.

where the operators ∇_r, ∇_c compute spatial differences along rows and columns, respectively. We call this regularization an ℓ_2 norm-based total variation (TV-ℓ_2) in this paper. Figure 2 illustrates the effect of the depth regularization. From Figure 2, we observe that smoothness regularization improves the loss function by removing several local minima. We also observed this effect in our simulations for high-dimensional depth recovery problem, which is not very sensitive to initialization with depth regularization.

Weighted ℓ_2 regularization on spatial gradients. Even though smoothness regularization on the inverse depth map removes some local minima and helps with converge, it does not respect the sharp edges in the depth map. To preserve sharp discontinuities in the (inverse) depth map, we used the following adaptive weighted regularization:

$$R_W(\boldsymbol{\alpha}) = \sum_{i,j=1}^{N} W^c_{i,j}(\alpha_{i,j} - \alpha_{i+1,j})^2 + W^r_{i,j}(\alpha_{i,j} - \alpha_{i,j+1})^2, \tag{16}$$

where $W^{r,\alpha}_{i,j}$ and $W^{c,\alpha}_{i,j}$ denote weights for row and column differences, respectively. We aim to select these weights to promote depth similarity for neighboring pixels, but avoid smoothing the sharp edges. To promote this, we selected weights with exponential decay in our experiments that we compute as

$$W^r_{i,j} = \exp\left(-\frac{(\alpha_{i,j} - \alpha_{i+1,j})^2}{\sigma}\right)$$
$$W^c_{i,j} = \exp\left(-\frac{(\alpha_{i,j} - \alpha_{i,j+1})^2}{\sigma}\right). \tag{17}$$

Such a weighted regularization forces pixels that have depth within a small range of one another to be smooth and does not penalize the points that have larger gap in depth (which indicates the presence of an edge). This helps preserve sharp edges in the reconstructed depth estimates. This weighting approach is analogous to bilateral filtering approach for image denoising [35], [36].

To highlight the effect of the weighted smoothness regularization on depth, we plot the following weighted quadratic function $f(\alpha_i - \alpha_j) = (\alpha_i - \alpha_j)^2 \exp(-(\alpha_i - \alpha_j)^2/\sigma)$ in Figure 3, where α_i, α_j stand for inverse depth of neighboring pixels. We plot the weighted function for different values of σ along with a normal quadratic function. The plots show that the quadratic function (without any weights) penalizes large values of depth differences; however, weighted function add small penalty if the neighboring pixels have large depth difference (which indicates the presence of an edge).

The regularized estimation problem for image and depth can be written in the following form:

$$\underset{\boldsymbol{\alpha},\mathbf{l}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2 + \lambda R_W(\boldsymbol{\alpha}). \tag{18}$$

We call this regularization approach weighted TV-ℓ_2 and solve it by alternately updating the inverse depth map $\boldsymbol{\alpha}$ and light intensity $\mathbf{l}$. A pseudocode of the algorithm is presented at Algorithm 1.

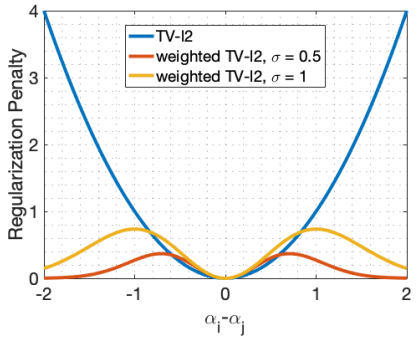


Fig. 3: The weighted regularization function penalizes depth values that are within a small distance of one another and does not penalize those values that are above certain threshold. The smooth range can be changed by tuning the parameter σ. In contrast to the TV-ℓ_2, a weighted TV-ℓ_2 regularization term does not penalize neighboring pixels with large depth disparity, which tends to preserve the sharpness of the edges in the depth estimation.

ℓ_1 regularization on spatial gradients. It is well-known that the ℓ_1 norm regularization enforces the solution to be sparse. We add an ℓ_1-based total variation norm [37] of the depth to our optimization problem. By enforcing the sparsity of spatial gradients, the edges of (inverse) depth map can be preserved. The ℓ_1 norm-based TV regularization term is given as

$$R_{TV}(\boldsymbol{\alpha}) = \sum_{i,j=1}^{N} |\alpha_{i,j} - \alpha_{i+1,j}| + |\alpha_{i,j} - \alpha_{i,j+1}|$$
$$= \|\nabla_r \boldsymbol{\alpha}\|_1 + \|\nabla_c \boldsymbol{\alpha}\|_1. \tag{19}$$

To solve the nonlinear optimization problem with ℓ_1 norm

Algorithm 1 Weighted TV-ℓ_2 regularized optimization

Input: Sensor measurements: $\mathbf{y}$
Output: Light distribution and inverse depth map: $\mathbf{l}, \boldsymbol{\alpha}$
Initialization via greedy algorithm:
Compute $\boldsymbol{\alpha}$ and $\mathbf{l}$ with depth pursuit algorithm in [5].
Refinement via alternating gradient descent:
for $k = 1 : k_{\max}$ **do**
$\widehat{\boldsymbol{\alpha}}^k = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}^{k-1}\|_2^2 + \lambda R_W(\boldsymbol{\alpha})$
$\widehat{\mathbf{l}}^k = \underset{\mathbf{l}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha}^k)\mathbf{l}\|_2^2$
end for
return $\widehat{\mathbf{l}}$ and $\widehat{\boldsymbol{\alpha}}$

regularization, we write the optimization problem as

$$\underset{\boldsymbol{\alpha},\mathbf{l}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2 + \lambda(\|\mathbf{d}_r\|_1 + \|\mathbf{d}_c\|_1)$$
$$\text{s.t. } \mathbf{d}_r = \nabla_r \boldsymbol{\alpha}, \ \mathbf{d}_c = \nabla_c \boldsymbol{\alpha}. \tag{20}$$

We solve this problem (20) using a split-Bregman method [38].

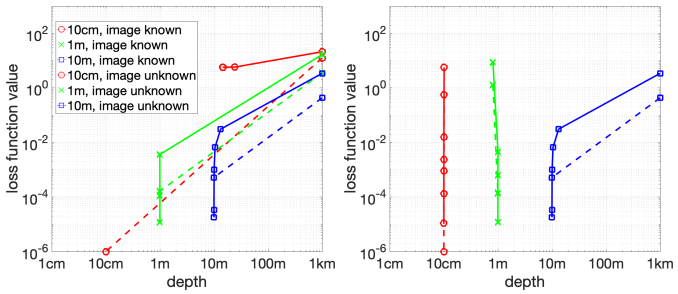
IV. Simulations and Results

A. Simulation Setup

To validate the performance of the proposed algorithm, we simulate a lensless imaging system using a binary planar mask with a separable maximum length sequence (MLS) pattern [39] that is placed 4mm away from a planar sensor array. We used an MLS sequence of length 1024 and converted all the -1s to 0s to create a separable binary pattern. We used square mask features, each of which is 30μm wide. Since we require the mask to be differential, we cannot use a binary mask pattern. Therefore, we convolved the binary pattern with a Gaussian blur kernel of length 15μm and standard deviation 5. The sensor contains 512×512 square pixels, each of which is 50μm wide. The chief ray angle of each sensor pixel is $\pm 18^\circ$. We assume that there is no noise added to the sensor measurements. In our experiments for continuous depth estimation, we fixed all the parameters to these default values and analyze the performance with respect to a single parameter.

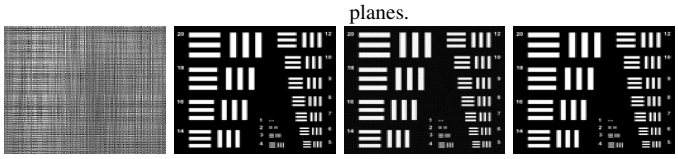
B. Reconstruction of a Single Plane

To understand various parameters of our method and their effect on depth estimation, we perform a simple simulation experiment in which the scene consists of a single plane and our goal is to recover the correct depth accurately. In other words, the depth map is parameterized by a scalar $\boldsymbol{\alpha}$ instead of a matrix of the same size as intensity $\mathbf{l}$. The estimation of the single depth parameter does not involve the regularization, so we solve (6) via alternating minimization using Algorithm 1 without the regularization term. We test the performance of our optimization program using a USAF target image placed at three different depths in the scene: $[10cm, 1m, 10m]$. We test two different initialization schemes for comparison. (1) Initial value of $\boldsymbol{\alpha}$ is selected to set depth at 1km (i.e., $\boldsymbol{\alpha} \approx 1$ for all the experiments). (2) We choose a depth plane from 10 candidate depths that were uniformly sampled in α to get effective depth range from 9cm to 1km; for a given measurement we selected $\boldsymbol{\alpha}$ that provided smallest loss function in (6). To separate the effect of image estimate and depth estimate, we tested two cases: one in which image intensity is an optimization variable and updated iteratively and the other one where the image intensity is fixed to its original value. Since the problem in (6) is nonconvex in $\boldsymbol{\alpha}$ even if we fix the value of intensity $\mathbf{l}$, there is no guarantee we can estimate the correct value of $\boldsymbol{\alpha}$.



(a) Convergence of estimates; initial depth at 1km.

(b) Convergence of estimates; initial depth selected from 10 candidate planes.



(c) Initial depth 1km. Image intensity unknown

(d) Initial depth 1km. Image intensity known

(e) Initial depth selected from 10 planes. Image intensity unknown

(f) Initial depth selected from 10 planes. Image intensity known

Fig. 4: Recovery of a single depth parameter. USAF target is tested at three different depths with two different types of initialization. (a) Initial depth is set to 1km. (b) Initial depth is selected out of 10 candidate planes. (a), (b) Reconstruction loss plotted against the depth estimate at every iteration of the algorithm. If the algorithm estimates correct image intensity and depth, the plot should converge to the true depth with a small loss function value. Solids lines correspond to the case when we jointly estimate image intensity and depth. Dashed lines correspond to the case when image intensities are known. (c)–(d) Reconstructed images when the original scene is 10cm away, under different choice of initial depth and unknown/known image intensities.

We report the results for 12 experiments (image at three different depths with two different initialization and known or unknown image intensity) in Figure 4. Figures 4(a) and (b) plot the loss function at the estimated depth at every iteration of the refinement step in Algorithm 1 for the cases when the initial depth is set to 1km and when the initial depth is selected out of 10 candidate depths. From the curves in Figure 4(a), we observe that we can estimate the depth accurately in all three cases when the image intensities are known (dashed lines). If we jointly estimate the depth and image intensities, the initial value of depth plays a critical role (solid lines). We observe that the intensity and depth are recovered correctly for the scene at 1m and 10m but the algorithm fails to recover correct depth for the scene at 10cm. This is mainly because

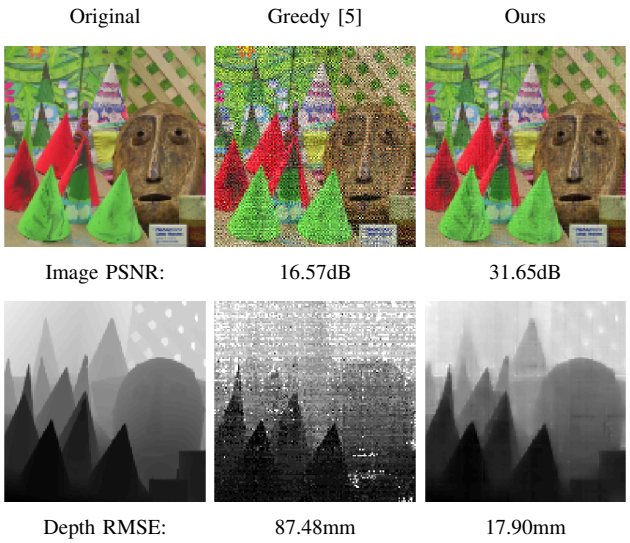


Fig. 5: Left to right: original image and depth of Cones scene; image and depth initialized via greedy algorithm [5]; depth estimation using weighted ℓ_2-based regularization. The depth in this scene varies from around 0.99m to 1.7m.

the initial value of $\boldsymbol{\alpha}$ is very far from the true value and the algorithm is more likely to get stuck in a local minima. Figure 4(c) shows the reconstructed image when the original scene is at 10cm away, initial depth is set to 1km, and we estimate image intensity and depth by solving (6). On the other hand, when we pick initial depth close to the true depth (out of 10 candidates, using a greedy approach), as shown in Figure 4(b), the algorithm converges to the true depth and recover the correct image intensities as well.

C. Reconstruction of Scenes with Continuous Depth

Depth datasets: We performed all our experiments on 3D images created using light intensities and depth information from Middlebury [40], Make3D [41], [42] and NYU Depth [43], the test scenes and their depth ranges are listed in Table I.

Test datasets	Min depth (m)	Max depth (m)
Sword	0.65	0.95
Cones	0.99	1.70
Playtable	1.47	3.75
Corner	3.93	10.60
Whiteboard	1.08	2.90
Playroom	1.62	2.93
Moebius	0.74	1.23
Books	0.73	1.27

TABLE I: Analysis experiments are performed on multiple scenes picked from Middlebury [40], Make3D [41], [42] and NYU Depth [43]. Results of the four scenes are presented within the main text, while the rest of them are reported in the supplemnetary material.

Initialization via greedy method: Let us further discuss our simulation setup using *cones* scene, for which the results are presented in Figure 5. We simulated the 3D scene using depth data from Middlebury dataset [40]. We sample the scene at uniform angles to create a 128×128 image and its (inverse) depth map with the same size. We can compute the physical depth from $\boldsymbol{\alpha}$ using (2). In our simulation, the depth of this scene ranges from around 0.99m to 1.7m. We used depth pursuit greedy algorithm in [5] as our initialization method. We selected 15 candidate depths by uniformly sampling the inverse depth values $\boldsymbol{\alpha}$ from 0.996 to 0.9976, which gives an effective depth in the same range as the original depth. Since we are trying to gauge the performance for off-the-grid estimate of depth, the candidate values of $\boldsymbol{\alpha}$ are not exactly the same as the true values of $\boldsymbol{\alpha}$ in our simulations. The output of initialization algorithm is then fed into the alternating gradient descent method.

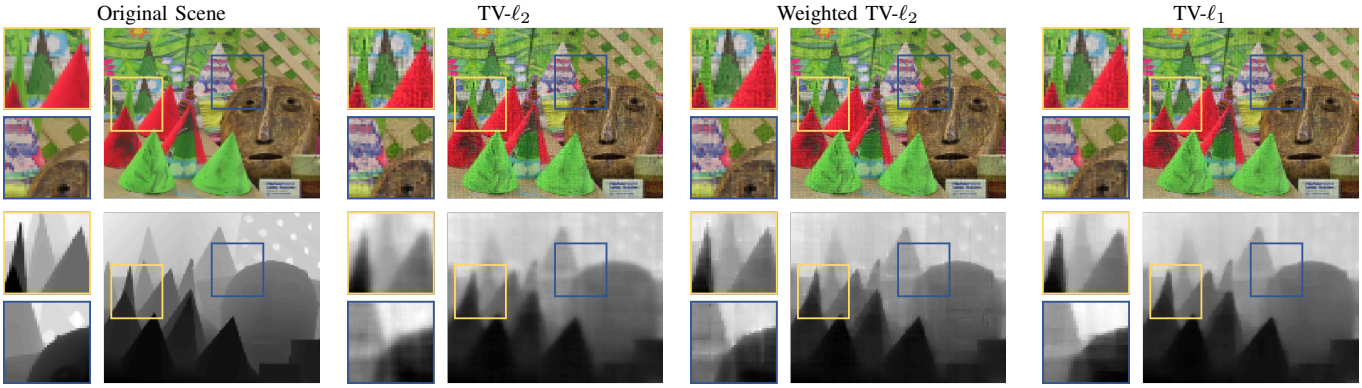
Performance metrics: We evaluate the performance of recovered image intensity and depth independent of each other. We report the peak signal to noise ratio (PSNR) of the estimated intensity images and root mean squared error (RMSE) of the estimated depth maps for all our experiments. The estimates for image intensity and depth maps for the initialization and our proposed weighted TV-ℓ_2 method are shown in Figure 5, along with the PSNR and RMSE. We can observe that both image and depth estimation from greedy method [5] contain several spikes because of the model mismatch with the predefined depth grid. In contrast, many of these spikes are removed in the estimations from the proposed algorithm with weighted TV-ℓ_2 while the edges are preserved.

Comparison of regularization methods: Here we present a comparison between three different regularization approaches. We reconstruct image intensity and (inverse) depth map using same measurements with TV-ℓ_2, weighted TV-ℓ_2, and TV-ℓ_1 regularization. The results are shown in Figure 6. Compared to the TV-ℓ_2 method, we observe that both weighted TV-ℓ_2 and TV-ℓ_1 preserve the sharp edges in image and depth estimates. Overall, in our experiments, weighted TV-ℓ_2 provided best results. Therefore, we used that as our default method in the rest of the paper.

D. Effects of Noise

Sensor noise exists widely in any observation process. The amplitude of noise depends on the intensities of sensor measurements and can adversely affect the reconstruction results. To investigate the effect of noise on our algorithm, we present simulation results for reconstruction of scenes from the same sensor measurements under different levels of additive white Gaussian noise. The experiments are performed on multiple 3D scenes listed in Table I. Some examples of reconstruction with different levels of noise are shown in Figure 7.

The plots recording PSNR of image intensities and RMSE of depth maps over a range of measurement SNR values are presented in Figure 8. As we can observe from the curves that the quality of both estimated image and depth improve when the measurements have small noise (high SNR) and the quality degrades as we add more noise in the measurements (low SNR). Another observation we can make is that the scenes that are farther away have higher RMSE. This aspect is understandable because as the scenes move farther, $\boldsymbol{\alpha}$ of the



(a) Image and depth of the original scene. The selected cones scene is taken from Middlebury dataset [40]. The range of depth is from 0.99 to 1.7 meters.

(b) Image and depth reconstruction from isotripic total variation. PSNR of image is 29.69dB and depth RMSE is 25.21mm.

(c) Image and depth reconstruction from weighted ℓ_2 total variation. The PSNR of image is 31.65dB and the RMSE of depth is 17.90mm. The edges of depth are preserved better.

(d) Image and depth reconstruction from TV-ℓ_1. The PSNR of image is 30.82dB and the depth RMSE is 19.56mm. The edges of depth are preserved better well.

Fig. 6: Comparison between reconstructions using three different regularization approaches from the same measurements.

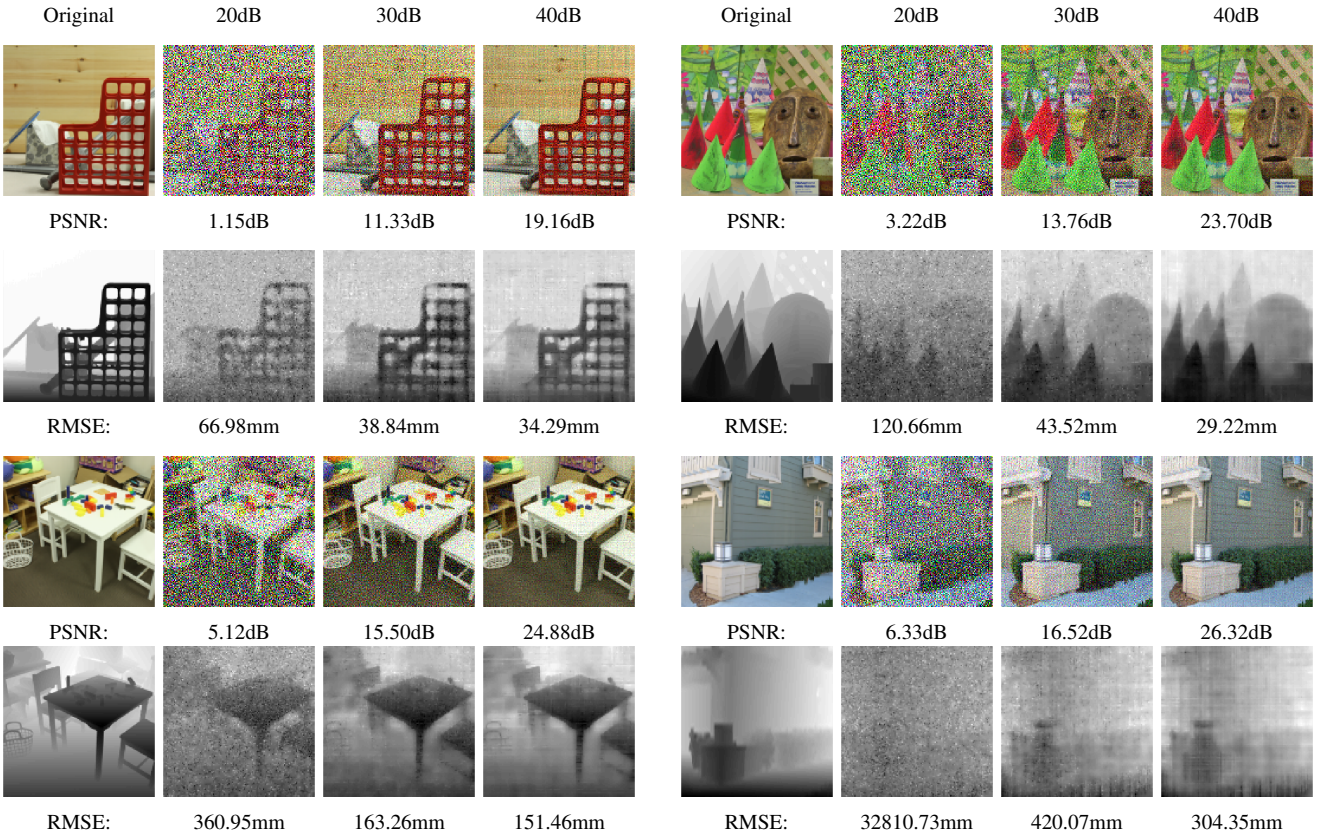
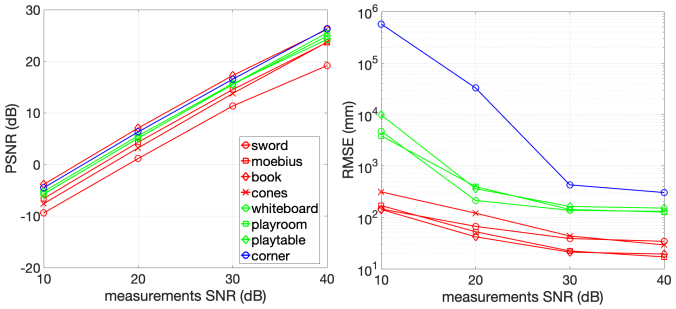


Fig. 7: Effects of noise: Reconstruction from the measurements with signal-to-noise ratio (SNR) at 20dB, 30dB and 40dB, along with the **PSNR** of reconstructed image and **RMSE** of reconstructed depth map. As expected, the quality of reconstructed image and depth improves as the noise level is reduced. The sequence in top left is for *Sword*, top right is *Cones*, bottom left *Playtable*, and bottom right is *Corner* scene from our test datasets.



(a) Image PSNR for different noise levels

(b) Depth RMSE for different noise levels

Fig. 8: Reconstruction from measurements with different levels of Gaussian noise on multiple scenes. Both of the image Peak Signal-Noise Ratio and depth Root mean squared error are improved as the noise is reduced. The reconstruction quality degrades if the scene is placed farther from the camera.

scene pixels all get very close to 1 and we cannot resolve fine depth variations in the scene.

E. Number of Sensor Measurements

In this experiment, we evaluate the performance of our algorithm as we increase/decrease the number of sensor measurements. The depth estimation problem we are solving is highly ill-posed because of existence of nontrivial null space of the system matrix and nonlinear dependence of measurements on the depth parameters. Adding more measurements helps with improve the solution of the system by adding more constraints on the feasible solutions.

We perform experiments with different number of sensor pixels while the size of each pixel is fixed as 50μm. We do not add any noise in these experiments to avoid randomness that potentially affect comparison. As we increase or decrease the number of pixels, it is equivalent to increasing or decreasing the sensor area. Therefore, when we increase the number of sensor pixels (equivalently, sensor area), the baseline of the sensor is also increased, which helps us in resolving the depth more accurately. The results are presented in Figure 9 for sensors of size 256×256, 512×512, and 1024×1024. We observe that the quality of both image and depth improves as we use more sensor pixels for measurements.

F. Size of Sensor

In conventional disparity-based depth estimation method [1], the quality of reconstructed depth depends on the disparity between frames captured from multiple camera views. Larger distance between camera viewing positions results in better depth estimation accuracy. In a lensless imaging system, we can think of each pinhole on the mask and the sensor area behind the mask as a tiny pinhole camera. The analogy only goes this far, because we do not record images from these tiny pinhole cameras separately; instead, we record a multiplexed version of the all the views. The disparity between different points on the sensors, however, does affect our ability to resolve depth of the scene, which is determined by the size of sensor.

To analyze the effect of disparity in our system, we performed experiments with three different sizes of sensor pixels from 25μm, 50μm, and 100μm. For a fair comparison, the number of sensor pixels and other parameters are set to the default settings as described earlier. No noise is included in this experiment. Results in terms of reconstructed image and depth maps are presented in Figure 10, where we observe that the quality of depth reconstruction improves as we increase the size of sensor pixels. The results in Figure 10 and 9 demonstrate that increasing the disparity of viewing points increases the depth reconstruction quality.

G. Comparison with Existing Methods

Finally, we present a comparison of our proposed algorithm and two other methods for 3D recovery with lensless cameras. In our method, we estimate light intensity and a depth map over continuous domain. The greedy method in [5] also estimates intensity and depth separately, but the depth map for any angle is restricted to one of the predetermined planes. Three-dimensional recovery using lensless cameras for 3D fluorescence microscopy was presented in [13] and [14], which estimate the entire 3D volume of the scene sampled over a predetermined 3D grid. Since the unknown volumetric scene in microscopy is often very sparse, the 3D scene recovery problem is solved as a sparse recovery problem for the light intensity over all the grid voxels. The result is a light distribution over the entire 3D space. We call this method 3D Grid and use the code provided in [13] to solve the 3D recovery problem using the forward model and measurements from our simulation setup.

The scenes studied in [14] and [13] are mostly transparent and contain multiple point light sources at different depths but the same angle. This is different from the natural scenes we are testing, where the objects are usually opaque and block light from objects behind them. We can model such scenes as having only one voxel along any angle to be nonzero; however, that will be a nonconvex constraint and to enforce that we will have to resort to some heuristic similar to the one in [5]. For the sake of comparison, we solve the ℓ_1 norm-based sparse recovery problem as described in [13], but then we pick the points with the maximum light intensity at each angle to form the reconstructed image and (inverse) depth map.

A comparison of different recovery methods with same imaging setup is shown in Figure 11. For the same scene, we reconstruct the same measurements using the three methods. As we can observe that our proposed algorithm offers a significant improvement compared to existing methods in all the test scenes.

V. Conclusion

We presented a new algorithm to jointly estimate the image and depth of a scene using a single snapshot of a mask-based lensless camera. Existing methods for 3D lensless imaging either estimate scene over a predefined 3D grid (which is computationally expensive) or a finite number of candidate

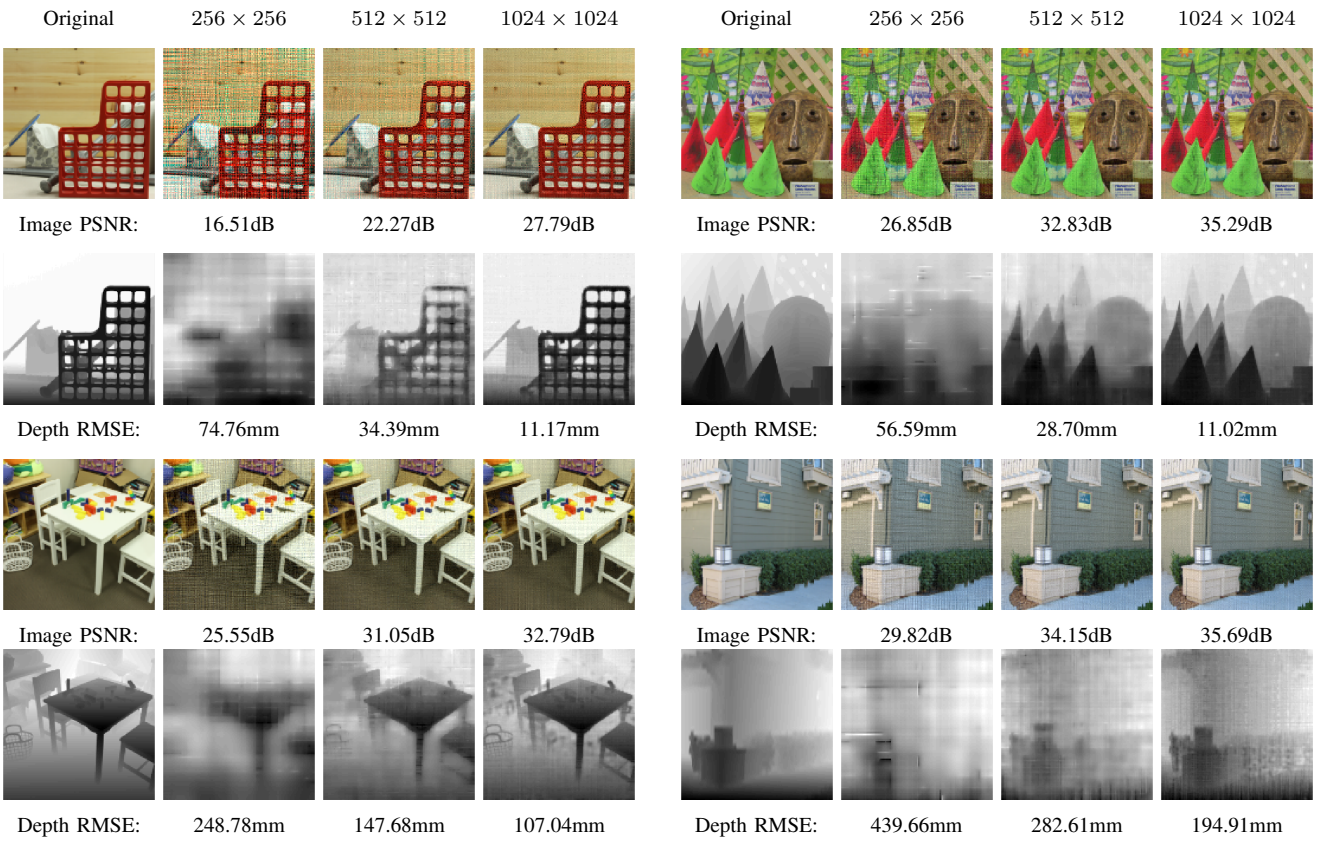


Fig. 9: Reconstructions from measurements with different numbers of pixels. The size of each sensor pixel is fixed as 50μm. The sequence in top left is for *Sword*, top right is *Cones*, bottom left *Playtable*, and bottom right is *Corner* dataset.

depth planes (which provides a coarse depth map). We divide the scene into an intensity map on uniform angles and a depth map on a continuous domain, which allows us to estimate a variety of scenes with different depth ranges using the same formulation. We jointly estimate the image intensity and depth map by solving a nonconvex problem. We initialize our estimates using a greedy method and add weighted regularization to enforce smoothness in the depth estimate while preserving the sharp edges. We demonstrated with extensive simulations that our proposed method can recover image and depth with high accuracy for a variety of scenes. We evaluated the performance of our methods under different noise levels, sensor sizes, and numbers of sensor pixels and found the method to be robust. Finally, we presented a comparison with existing methods for lensless 3D imaging and demonstrated that our method provides significantly better results.

REFERENCES

[1] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[2] S. B. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor - system description, issues and solutions," in *Conference on Computer Vision and Pattern Recognition Workshop*, June 2004, pp. 35–35.

[3] Felix Heide, Matthias B Hullin, James Gregson, and Wolfgang Heidrich, "Low-budget transient imaging using photonic mixer devices," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, pp. 45, 2013.

[4] M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "Flatcam: Thin, lensless cameras using coded aperture and computation," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 384–397, Sept 2017.

[5] M Salman Asif, "Lensless 3d imaging using mask-based cameras," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2018, pp. 6498–6502.

[6] Yan Liu, Jianhua Ma, Yi Fan, and Zhengrong Liang, "Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction," *Physics in Medicine & Biology*, vol. 57, no. 23, pp. 7923, 2012.

[7] Adam Yedidia, Christos Thrampoulidis, and Gregory Wornell, "Analysis and optimization of aperture design in computational imaging," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4029–4033, April 2018.

[8] E. E. Fenimore and T. M. Cannon, "Coded aperture imaging with uniformly redundant arrays," *Appl. Opt.*, vol. 17, no. 3, pp. 337–347, Feb 1978.

[9] A. Busboom, H. Elders-Boll, and H. D. Schotten, "Uniformly redundant arrays," *Experimental Astronomy*, vol. 8, no. 2, pp. 97–123, Jun 1998.

[10] T. M. Cannon and E. E. Fenimore, "Coded Aperture Imaging: Many Holes Make Light Work," *Optical Engineering*, vol. 19, pp. 283, June 1980.

[11] Vivek Boominathan, Jesse K Adams, M Salman Asif, Benjamin W Avants, Jacob T Robinson, Richard G Baraniuk, Aswin C Sankaranarayanan, and Ashok Veeraraghavan, "Lensless imaging: A computational renaissance," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 23–35, 2016.

[12] G. D. DeMeester H. Scharfman H. H. Barrett, D. T. Wilson, "Fresnel zone plate imaging in radiology and nuclear medicine," *Optical Engineering*, vol. 12, no. 1, pp. 8 – 12 – 5, 1973.

[13] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller, "Diffusercam: lensless single-exposure 3d imaging," *Optica*, vol. 5, no. 1, pp. 1–9, Jan 2018.

[14] Jesse K. Adams, Vivek Boominathan, Benjamin W. Avants, Daniel G. Vercosa, Fan Ye, Richard G. Baraniuk, Jacob T. Robinson, and Ashok Veeraraghavan, "Single-frame 3d fluorescence microscopy with ultra-miniature lensless flatscope," *Science Advances*, vol. 3, no. 12, 2017.

[15] Z. Tan, P. Yang, and A. Nehorai, "Joint sparse recovery method for compressed sensing with structured dictionary mismatches," *IEEE*

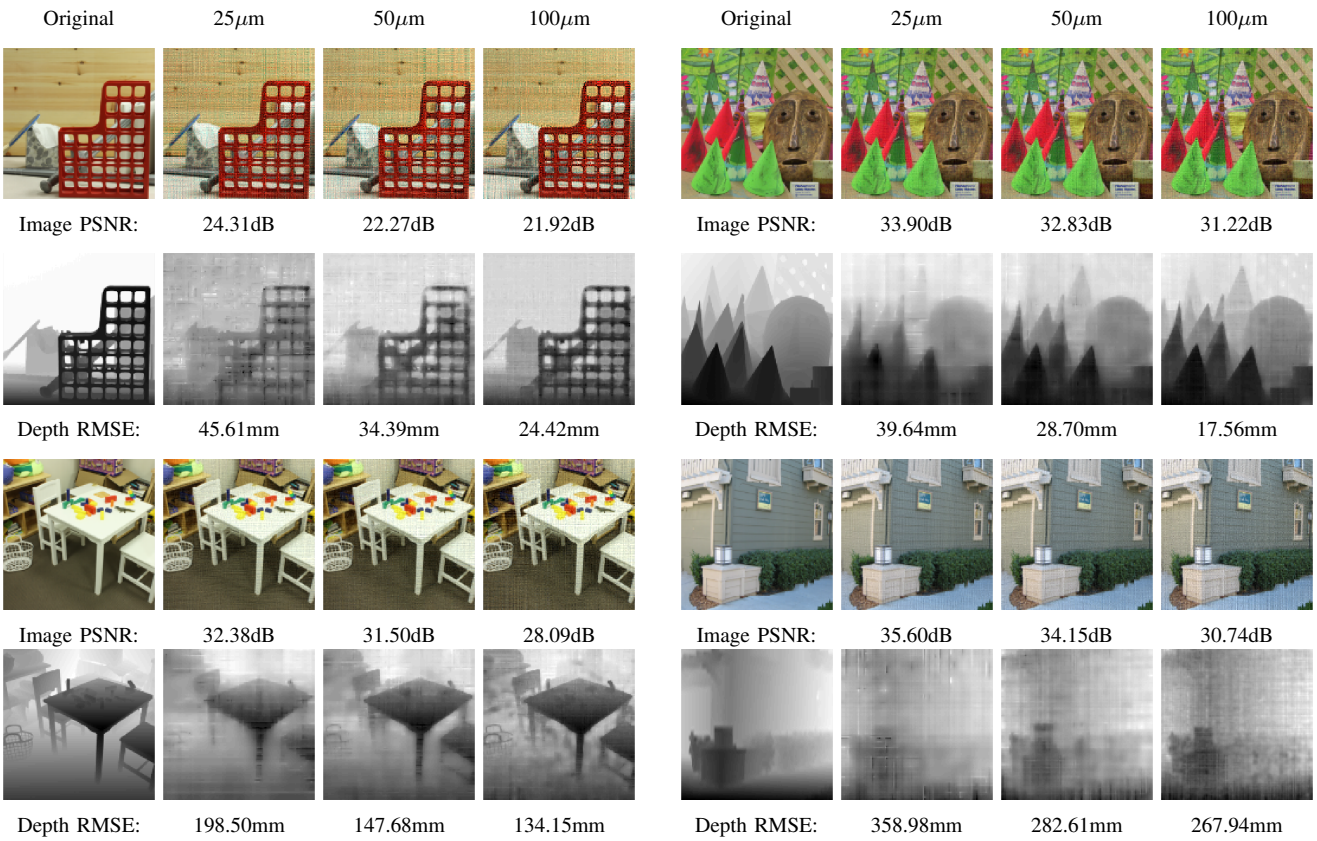


Fig. 10: Reconstructions from measurements with different sizes of sensor pixels. The number of sensor pixels is fixed as 512×512. The quality of depth reconstruction improves as we increase the size of sensor pixels.

Transactions on Signal Processing, vol. 62, no. 19, pp. 4997–5008, Oct 2014.

[16] N. Boyd, G. Schiebinger, and B. Recht, "The alternating descent conditional gradient method for sparse inverse problems," in *IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec 2015, pp. 57–60.

[17] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, Dec 2007.

[18] Deanna Needell and Joel A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Commun. ACM*, vol. 53, no. 12, pp. 93–100, Dec. 2010.

[19] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[20] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, Dec 2012.

[21] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7465–7490, Nov 2013.

[22] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse bayesian inference," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 38–43, Jan 2013.

[23] Dharmpal Takhar, Jason N. Laska, Michael B. Wakin, Marco F. Duarte, Dror Baron, Shriram Sarvotham, Kevin F. Kelly, and Richard G. Baraniuk, "A new compressive imaging camera architecture using optical-domain compression," *Proc.SPIE*, vol. 6065, pp. 6065 – 6065 – 10, 2006.

[24] A. Zomet and S. K. Nayar, "Lensless imaging with a controllable aperture," in *IEEE Computer Vision and Pattern Recognition*, June 2006, vol. 1, pp. 339–346.

[25] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, March 2008.

[26] D. Reddy, J. Bai, and R. Ramamoorthi, "External mask based depth and light field camera," in *2013 IEEE International Conference on Computer Vision Workshops*, Dec 2013, pp. 37–44.

[27] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. Graph.*, vol. 26, no. 3, July 2007.

[28] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," in *ACM transactions on graphics (TOG)*. ACM, 2007, vol. 26, p. 69.

[29] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d learning phase masks for passive single view depth estimation," in *2019 IEEE International Conference on Computational Photography (ICCP)*, May 2019, pp. 1–12.

[30] Julie Chang and Gordon Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proc. IEEE ICCV*, 2019.

[31] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.

[32] D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–14, Jan 2020.

[33] Mark Schmidt, "minfunc: unconstrained differentiable multivariate optimization in matlab," 2005.

[34] Dong C. Liu and Jorge Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1, pp. 503–528, Aug 1989.

[35] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Jan 1998, pp. 839–846.

[36] Frédo Durand and Julie Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, July 2002.

[37] Leonid I. Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total

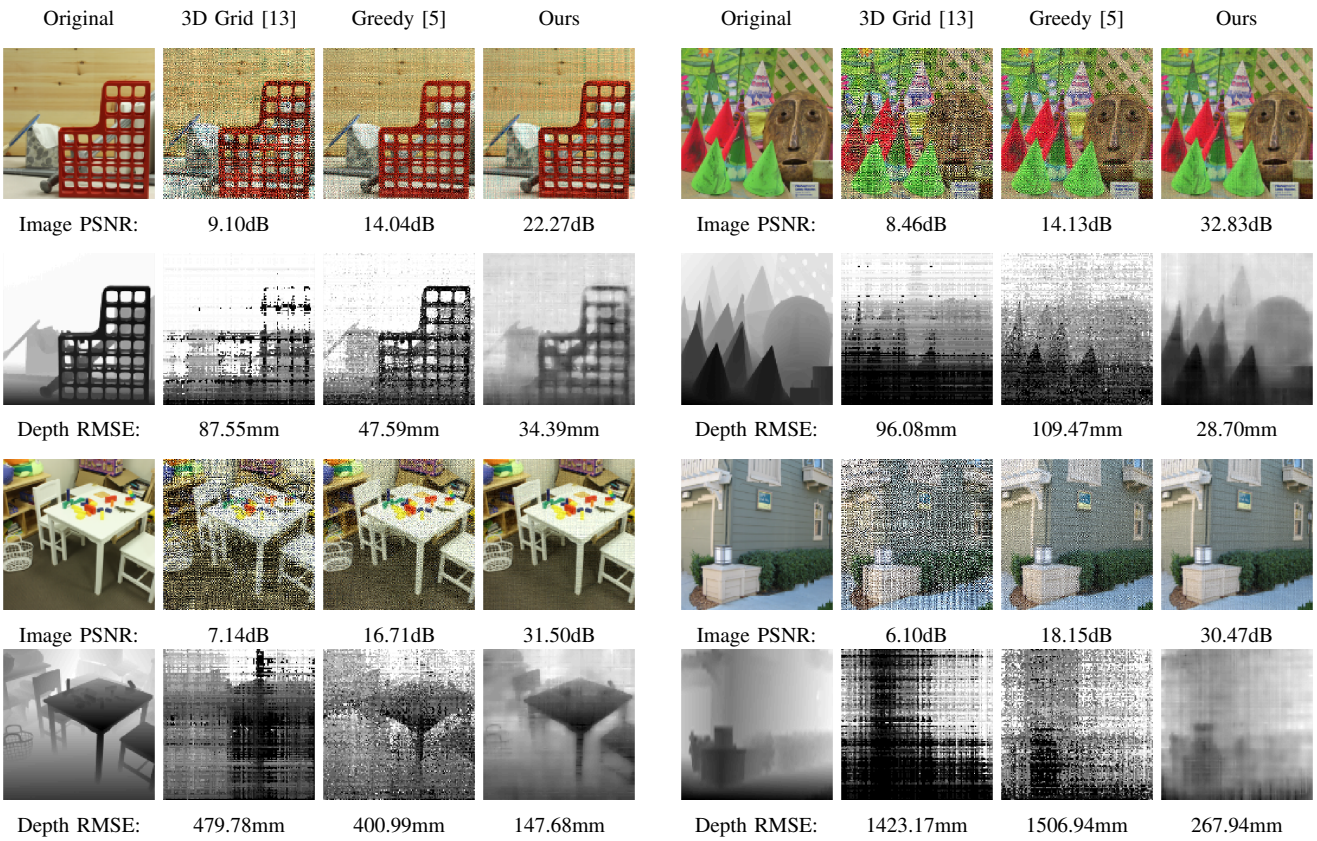


Fig. 11: Comparison of existing 3D recovery methods for lensless imaging, 3D grid method in [13], [14] and greedy method in [5], with our proposed method. 3D grid method provides a 3D volume with multiple depth planes; therefore, we pick the depth with the largest light intensity along any angle for comparison.

variation based noise removal algorithms," *Phys. D*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992.

[38] Tom Goldstein and Stanley Osher, "The split bregman method for l1-regularized problems," *SIAM J. Img. Sci.*, vol. 2, no. 2, pp. 323–343, Apr. 2009.

[39] F. J. MacWilliams and N. J. A. Sloane, "Pseudo-random sequences and arrays," *Proceedings of the IEEE*, vol. 64, no. 12, pp. 1715–1729, Dec 1976.

[40] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision*, Dec 2001, pp. 131–140.

[41] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng, "Learning depth from single monocular images," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds., pp. 1161–1168. MIT Press, 2006.

[42] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, May 2009.

[43] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.